Combinatorial generalization bounds[*]

K. V. Vorontsov
Computing Center RAS
voron@forecsys.ru

A. A. Ivahnenko
Moscow Institute
of Physics and Technology
ivahnenko@forecsys.ru

P. V. Botov
Moscow Institute
of Physics and Technology
pbotov@forecsys.ru

I. M. Reshetnyak
Moscow State University
ilya.reshetnyak@gmail.com

I. O. Tolstikhin
Computing Center RAS
iliya.tolstikhin@gmail.com

Abstract

In this paper we propose a new combinatorial technique for obtaining data dependent generalization bounds. We introduce a splitting and connectivity graph (SC-graph) over the set of classifiers. In some cases the knowledge of this graph leads to an exact generalization bound. Typically, the knowledge of a little part of the SC-graph is sufficient for reasonable approximation of the bound. Being applied to a parametric set of conjunctive rules our bound helps to obtain more reliable classifiers as compositions of less overfitted rules.

1 Introduction

The accurate bounding of overfitting remains one of the most challenging open problems in computational learning theory starting with VC-theory (Vapnik and Chervonenkis, 1971). Generalization bounds are still pessimistically overestimated regardless of recent significant improvements, see surveys (Vayatis and Azencott, 1999, Langford, 2002, Boucheron et al., 2005). Conservative bounds are not always suitable for overfitting understanding, prediction, and control. Another difficulty is that both final and intermediate bounds are usually expressed in terms of unobservable quantities. Therefore it is not always possible to measure and compare the factors of overestimation. For classical VC bounds such empirical measurement has been performed in (Vorontsov, 2008). The permutational probability framework has been developed to make intermediate bounds measurable. It has been shown that overfitting depends not only on the number of different classifiers (shattering coefficient) but in greater degree on their diversity. Splitting and connectivity properties of the set of classifiers are two aspects of diversity that reduce overfitting and might help to obtain tighter bounds.

The *splitting* property means that only a small part of classifiers from a given set have a low error rate and, as a result, a high chance to be produced by a learning algorithm. Splitting happens when we deal with a certain problem having a fixed target function.

The *connectivity* property means that for any classifier from a given parametric set there exist a number of classifiers from the set differing from the first one by exactly one object. Such classifiers are called *connected*. Connectivity owes to parameter continuity of a classification function.

Experiments with split and non-split chains of classifiers (Vorontsov, 2009) showed that the absence of one of these advantageous properties can result in significant overfitting even though the set contains a few dozens of classifiers. Hence, to possess both splitting and connectivity properties is a must for sets containing billions of classifiers in practice.

In this work we develop a combinatorial theory that accurately deals with both splitting and connectivity and gives tight or even exact bounds on probability of overfitting. Basic definitions and notations are introduced in section 2. Section 3 revisits two classical bounds in terms of permutational probabilities. In section 4 we introduce the principle of protective and prohibitive subsets fundamental for further considerations. Section 5 contains main theorems about splitting and connectivity (SC) bounds. In section 6 we show that SC-bounds can be exact for some nontrivial sets of classifiers. In section 7 then SC-bound is applied to the set of conjunctive rules. The overfitting reduction strategy is proposed that can be easily incorporated into existing rule induction engines. The experiment shows that the usage of SC-bound results in better generalization on real data sets.

[*] Supported by Russian Foundation for Basic Research grants 08-07-00422, 11-07-00480 and the program "Algebraical and combinatorial methods of cybernetics and new generation information systems" of Russian Academy of Sciences, Branch of Mathematical Sciences.

2 Definitions and notation

Let $\mathbb{X} = \{x_1, \dots, x_L\}$ be a set of objects and A be a set of classifiers. By $I\colon A \times X \to \{0, 1\}$ denote a binary loss function such that $I(a, x) = 1$ if a classifier a produces an error on an object x. For further considerations there is no need to specify what is "classifier". Particularly, regression function can also be a "classifier" if a binary loss function is used.

A binary vector $(a_i) \equiv \bigl(I(a, x_i)\bigr)_{i=1}^{L}$ of size L is called an *error vector* of the classifier a. Assume that all classifiers from A have pairwise different error vectors. The number of errors of a classifier a on a sample $X \subseteq \mathbb{X}$ is defined as $n(a, X) = \sum_{x \in X} I(a, X)$. Denote $n(a) = n(a, \mathbb{X})$ for short. The error rate is defined as $\nu(a, X) = \frac{1}{|X|} n(a, X)$. A *learning algorithm* is a mapping $\mu\colon 2^{\mathbb{X}} \to A$ that takes a training sample $X \subseteq \mathbb{X}$ and gives a classifier $\mu X \in A$.

Permutational (transductive) probability. By $[\mathbb{X}]^\ell$ denote a set of all $\binom{L}{\ell} = \frac{L!}{\ell!(L-\ell)!}$ samples $X \subset \mathbb{X}$ of size ℓ. Assume that all partitions of the set $\mathbb{X}$ into an observed training sample X of size ℓ and a hidden test sample $\bar{X} = \mathbb{X} \setminus X$ of size $k = L - \ell$ can occur with equal probability.

If the *discrepancy* $\delta(a, X) = \nu(a, \bar{X}) - \nu(a, X)$ is greater than a given nonnegative threshold ε, then the classifier $a = \mu X$ is said to be *overfitted*. Our goal is to estimate the *probability of overfitting*:

$$Q_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\bigl[\delta(\mu, X) \geqslant \varepsilon\bigr] = \frac{1}{\binom{L}{\ell}} \sum_{X \in [\mathbb{X}]^\ell} \bigl[\delta(\mu, X) \geqslant \varepsilon\bigr].$$

where $\delta(\mu, X) = \delta(\mu X, X)$ for short; the square brackets denote a transformation of a logical value into numerical one according to Iverson's convention: $[\mathit{true}] = 1$, $[\mathit{false}] = 0$ (Graham et al., 1994).

The *inversion* of an upper bound $Q_\varepsilon \leqslant \eta(\varepsilon)$ is an inequality $\nu(\mu X, \bar{X}) \leqslant \nu(\mu X, X) + \varepsilon(\eta)$ that holds with probability at least $1 - \eta$, where $\varepsilon(\eta)$ is the inverse function for $\eta(\varepsilon)$.

The permutational probabilistic framework seems to be restrictive because the fundamental notion of "probability" degenerates into the trivial "fraction of partitions". Nevertheless, the independence assumption is actually kept in this framework. This is quite sufficient to derive the law of large numbers, rank tests, VC-bounds, and many other fundamental statistical facts.

The permutational framework has three advantages:
1) it makes redundant some usual intermediate bounds as symmetrization (Philips, 2005);
2) it allows to measure probabilities empirically in a manner of cross-validation;
3) it encourages us to keep bounds in exact, non-asymptotical form.

Learning algorithms. *Empirical risk minimization* (ERM) is a classical and perhaps most natural example of the learning algorithm:

$$\mu X \in A(X) = \operatorname{Arg}\min_{a \in A} n(a, X). \tag{1}$$

The choice of a classifier minimizing empirical risk may be ambiguous because of the discreteness of the function $n(a, X)$. ERM algorithm μ is said to be *pessimistic* or *optimistic* if, respectively,

$$\mu X = \arg \max_{a \in A(X)} n(a, \bar{X}); \qquad \mu X = \arg \min_{a \in A(X)} n(a, \bar{X}).$$

The *probability* $\widetilde{Q}_\varepsilon$ *of a large uniform deviation* is considered as a main functional to be bounded in Vapnik-Chervonenkis and Rademacher Complexity theories of generalization:

$$Q_\varepsilon(\mu, \mathbb{X}) \leqslant \widetilde{Q}_\varepsilon(A, \mathbb{X}) \equiv \mathsf{P}\Bigl[\max_{a \in A} \delta(a, X) \geqslant \varepsilon\Bigr]. \tag{2}$$

The uniform functional $\widetilde{Q}_\varepsilon$ can give highly overestimated upper bounds on Q_ε. Nevertheless, it is widely used because its bounds hold for any learning algorithm μ. Minimizing the inversion of a bound leads to a new learning algorithm that tends to reduce overfitting, by construction. This is a standard way to get practical outcome from generalization bounds. A number of complexity penalization methods has been obtained in this way (Vapnik, 1998, Langford, 2002).

The uniform functional $\widetilde{Q}_\varepsilon$ can also be represented as a probability of overfitting for a special learning algorithm called *discrepancy maximization*:

$$\mu X = \arg \max_{a \in A} \delta(a, X).$$

The pessimistic ERM, optimistic ERM, and discrepancy maximization cannot be implemented in practice because they look into a hidden part of data $\bar{X}$ unknown at the learning stage. Nevertheless, they are very useful for theoretical considerations because of the following relationships.

Lemma 1 *For any set $\mathbb{X}$ and any ERM learning algorithm μ*

$$Q_\varepsilon(\mathrm{OptERM}, \mathbb{X}) \leqslant Q_\varepsilon(\mu, \mathbb{X}) \leqslant Q_\varepsilon(\mathrm{PessERM}, \mathbb{X}) \leqslant Q_\varepsilon(\mathrm{DiscrMax}, \mathbb{X}) = \widetilde{Q}_\varepsilon(A, \mathbb{X}).$$

3 Fixed classifier bound and Vapnik-Chervonenkis bound

In this section we show that permutational probabilistic framework gives an exact bound for a fixed classifier (FC) and an upper Vapnik-Chervonenkis (VC) bound by a quite straightforward way.

Hypergeometric distribution. For a classifier a such that $m = n(a, \mathbb{X})$ the probability to have s errors on a sample X is given by a hypergeometric function:

$$\mathsf{P}\bigl[n(a, X) = s\bigr] = \mathsf{P}\bigl[n(a, \bar{X}) = m - s\bigr] = h_L^{\ell,\, m}(s),$$

where $h_L^{\ell,\, m}(s) = \binom{m}{s}\binom{L-m}{\ell-s}/\binom{L}{\ell}$, argument s runs from $s_0 = \max\{0, m-k\}$ to $s_1 = \min\{m, \ell\}$, parameter m takes values $0, \dots, L$. It is assumed that $\binom{m}{s} = h_L^{\ell,\, m}(s) = 0$ for all other integers m, s.

Define the cumulative distribution function (left tail) of the hypergeometric distribution

$$H_L^{\ell,\, m}(z) = \sum_{s=s_0}^{\lfloor z \rfloor} h_L^{\ell,\, m}(s).$$

Consider a set $A = \{a\}$ containing only a fixed classifier, so that $\mu X = a$ for any X. Then the probability of overfitting Q_ε transforms into the probability of large deviation between error rates on two samples $X, \bar{X}$. If the number of errors $n(a)$ is known, then an exact Q_ε bound can be obtained.

Theorem 2 (FC-bound) *For any set $\mathbb{X}$, any $\varepsilon \in [0, 1]$, and a fixed classifier a such that $m = n(a)$ the probability of overfitting is given by the left tail of the hypergeometric distribution:*

$$Q_\varepsilon(a, \mathbb{X}) = H_L^{\ell,\, m}\bigl(\tfrac{\ell}{L}(m - \varepsilon k)\bigr). \tag{3}$$

Proof: Denote $s = n(a, X)$ and rewrite the overfitting condition $\delta(a, X) \geqslant \varepsilon$ as $\frac{1}{k}(m - s) - \frac{1}{\ell}s \geqslant \varepsilon$ or equivalently $s \leqslant \frac{\ell}{L}(m - \varepsilon k) \equiv s_m(\varepsilon)$. Then

$$Q_\varepsilon = \mathsf{P}\bigl[n(a, X) \leqslant s_m(\varepsilon)\bigr] = \sum_{s=s_0}^{\lfloor s_m(\varepsilon) \rfloor} \mathsf{P}\bigl[n(a, X) = s\bigr] = \sum_{s=s_0}^{\lfloor s_m(\varepsilon) \rfloor} h_L^{\ell,\, m}(s) = H_L^{\ell,\, m}(s_m(\varepsilon)).$$

■

The hypergeometric distribution plays a fundamental role in all further combinatorial bounds.

Theorem 3 (VC-bound) *For any set $\mathbb{X}$, any learning algorithm μ, and any $\varepsilon \in [0, 1]$ the probability of large uniform deviation is bounded by the sum of FC-bounds over the set A:*

$$\widetilde{Q}_\varepsilon(A, \mathbb{X}) \leqslant \sum_{a \in A} H_L^{\ell,\, m}\bigl(\tfrac{\ell}{L}(m - \varepsilon k)\bigr), \quad m = n(a). \tag{4}$$

Proof: Apply a union bound substituting the maximum of binary values by their sum:

$$\widetilde{Q}_\varepsilon = \mathsf{P} \max_{a \in A}\bigl[\delta(a, X) \geqslant \varepsilon\bigr] \leqslant \sum_{a \in A} \mathsf{P}\bigl[\delta(a, X) \geqslant \varepsilon\bigr] = \sum_{a \in A} H_L^{\ell,\, m}(s_m(\varepsilon)), \quad m = n(a).$$

■

The union bound is the only reason of the looseness of the VC-bound (4).

Note that further weakening gives a well known form of the VC-bound (Vapnik, 1998):

$$\widetilde{Q}_\varepsilon(A, \mathbb{X}) \leqslant |A| \max_m H_L^{\ell,\, m}(s_m(\varepsilon)) \leqslant |A| \cdot \tfrac{3}{2} e^{-\varepsilon^2 \ell}, \quad \text{if } \ell = k,$$

where $|A|$ is called a *shattering coefficient* of the set of classifiers A on the set $\mathbb{X}$.

It is well known that VC-bound is highly overestimated which can be explained by the fact that all classifiers make approximately equal contributions to the VC-bound. However, the set of classifiers is usually split into error rates in quite nonuniform manner. Most classifiers are bad, therefore, have vanishing probability to be obtained as a result of learning and make a negligible contribution to the probability of overfitting. On the other hand, similar classifiers share their contribution, thus each of them contributes poorly again. VC bound totally ignores these advantageous effects. The uniform deviation bound sacrifices the splitting effect; the union bound sacrifices the similarity effect.

Note that the initial proof by Vapnik and Chervonenkis was purely combinatorial. Later a combinatorial approach was neglected in favor of more sophisticated techniques from functional analysis and concentration of measure (Lugosi, 2003, Boucheron et al., 2000). We wish revisit this point and demonstrate that the combinatorial approach has not exhausted its potential.

4 The principle of protective and prohibitive subsets

The principle of protective and prohibitive sets (Vorontsov, 2010) is based on the conjecture that the necessary and sufficient condition for $\mu X = a$ can be specified explicitly for any classifier $a \in A$ in terms of subsets of objects. From this conjecture an exact Q_ε bound has been derived.

In this work we use a similar conjecture relaxed to the necessary condition and derive an upper bound which has a simpler form.

Conjecture 4 *For each classifier $a \in A$ there exists a protective subset $X_a \in \mathbb{X}$ and a prohibitive subset $X'_a \in \mathbb{X}$ such that for any $X \in [\mathbb{X}]^\ell$*

$$\big[\mu X = a\big] \leqslant \big[X_a \subseteq X\big]\big[X'_a \subseteq \bar{X}\big]. \tag{5}$$

The subset $\mathbb{X}\backslash X_a \backslash X'_a$ is called *neutral* for a classifier a. The presence or absence of neutral objects in a training sample X does not change the result of learning μX. Later we will give nontrivial examples of μ and A that satisfy conjecture 4.

Lemma 5 *If conjecture 4 holds, then the probability to learn a classifier a can be bounded:*

$$\mathsf{P}\big[\mu X = a\big] \leqslant P_a \equiv \tbinom{L_a}{\ell_a} / \tbinom{L}{\ell},$$

where $L_a = L - |X_a| - |X'_a|$ and $\ell_a = \ell - |X_a|$ are the number of neutral objects for a classifier a in the general set $\mathbb{X}$ and sample X respectively.

Proof: According to the conjecture $\mathsf{P}\big[\mu X = a\big] \leqslant \mathsf{P}\big[X_a \subseteq X\big]\big[X'_a \subseteq \bar{X}\big]$. The right-hand side is a fraction of partitions $\mathbb{X} = X \sqcup \bar{X}$ such that $X_a \subseteq X$ and $X'_a \subseteq \bar{X}$. The number of such partitions is equal to $\tbinom{L_a}{\ell_a}$. The number of all partitions is equal to $\tbinom{L}{\ell}$, hence their ratio gives P_a. ■

Theorem 6 *If conjecture 4 holds, then for any $\varepsilon \in [0,1]$ the bound on probability of overfitting is*

$$Q_\varepsilon \leqslant \sum_{a \in A} P_a H_{L_a}^{\ell_a,\, m_a}\left(s_a(\varepsilon)\right), \tag{6}$$

where $m_a = n(a, \mathbb{X}\backslash X_a \backslash X'_a)$ is a number of errors that classifier a produces on neutral objects and $s_a(\varepsilon) = \frac{\ell}{L}\big(n(a) - \varepsilon k\big) - n(a, X_a)$ is a largest number of errors $n(a, X\backslash X_a)$ that classifier a produces on neutral training objects provided that discrepancy $\delta(a, X)$ exceeds ε.

Proof: The probability of overfitting Q_ε can be found as a total probability from probability to learn each of classifiers $\mathsf{P}\big[\mu X = a\big]$ and conditional probabilities $Q_{\varepsilon|a} = \mathsf{P}\big[\delta(a, X) \geqslant \varepsilon \mid a = \mu X\big]$:

$$Q_\varepsilon = \sum_{a \in A} \mathsf{P}\big[\mu X = a\big] Q_{\varepsilon|a} \leqslant \sum_{a \in A} P_a Q_{\varepsilon|a}.$$

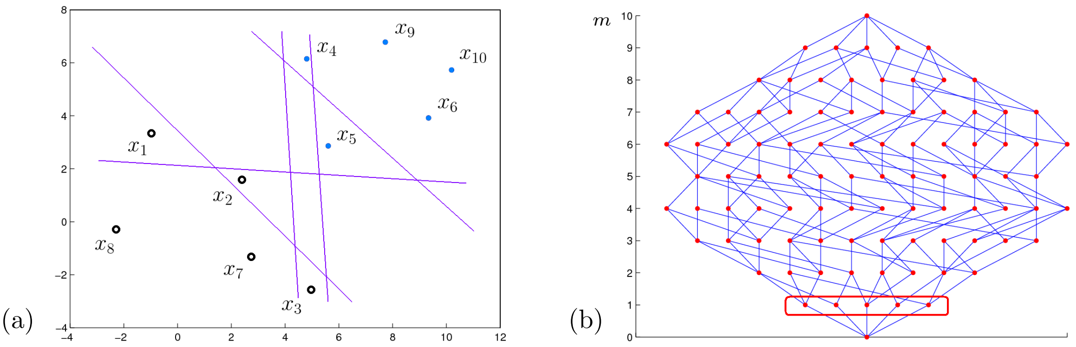
The conditional probability $Q_{\varepsilon|a}$ can be obtained from theorem 3 by taking into account that the subsets X_a and X'_a can not be involved in partitioning given a fixed classifier a. Only L_a neutral objects are partitioned into ℓ_a training and $L_a - \ell_a$ testing objects. To employ theorem 3 we express the discrepancy $\delta(a, X)$ in terms or the number of errors on neutral training objects $s = n(a, X\backslash X_a)$:

$$\delta(a, X) = \tfrac{1}{k}\big(n(a) - s - n(a, X_a)\big) - \tfrac{1}{\ell}\big(s + n(a, X_a)\big).$$

Condition $\delta(a, X) \geqslant \varepsilon$ is equivalent to $s \leqslant s_a(\varepsilon)$. Then $Q_{\varepsilon|a} = H_{L_a}^{\ell_a,\, m_a}\left(s_a(\varepsilon)\right)$ and (6) holds. ■

Note that the sum $\sum_a P_a$ can be interpreted as a degree of looseness of the bound (6). The bound is exact if this sum is equal to 1.

The principle of protective and prohibitive subsets is a powerful tool to obtain combinatorial generalization bounds. In (Vorontsov, 2010) it has been used to obtain exact bounds on probability of overfitting for model sets of classifiers like monotonic and unimodal chains. This work is focused on common bounds for arbitrary sets of classifiers that take into account both splitting and connectivity properties of the set.



(c)

	layer 0	layer 1					layer 2								
x_1	0	1	0	0	0	0	1	0	0	0	0	1	1	0	...
x_2	0	0	1	0	0	0	1	1	0	0	0	0	0	0	...
x_3	0	0	0	1	0	0	0	1	1	0	0	0	0	1	...
x_4	0	0	0	0	1	0	0	0	1	1	0	0	0	0	...
x_5	0	0	0	0	0	1	0	0	0	1	1	1	0	0	...
x_6	0	0	0	0	0	0	0	0	0	0	1	0	1	0	...
x_7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	...
x_8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
x_9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
x_{10}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...

Figure 1: Two-dimensional linearly separable classification task with $L = 10$ objects of 2 classes and 5 linear classifiers that produce exactly one error (a). The SC-graph over the set of all 2-dimensional linear classifiers (b). The first layer ($m = 1$) corresponds to 5 classifiers shown at the left chart. The fragment of error matrix corresponding to layers $m = 0, 1, 2$ (c).

5 Splitting and connectivity bounds

The splitting and connectivity graph. Define an order relation on classifiers $a \leqslant b$ as a natural order over their error vectors: $a_i \leqslant b_i$ for all $i = 1, \dots, L$. Define a metric on classifiers as a Hamming distance between error vectors: $\rho(a, b) = \sum_{i=1}^{L} |a_i - b_i|$. Classifiers a and b are called *connected* if $\rho(a, b) = 1$. Define the precedence relation on classifiers $a \prec b$ as $\bigl(a \leqslant b\bigr) \wedge \bigl(\rho(a, b) = 1\bigr)$.

The set of classifiers A can be represented by a multipartite directed graph $\langle A, E\rangle$ that we call the *splitting and connectivity graph* (SC-graph) in which vertices are classifiers, and edges (a, b) are pairs of classifiers such that $a \prec b$, see example on Figure 1. The partite subsets $A_m = \{a \in A \colon n(a) = m\}$ are called *error layers*, $m = 0, \dots, L$. Each edge of the SC-graph (a, b) corresponds to an object $x_{ab} \in \mathbb{X}$ such that $I(a, x_{ab}) = 0$ and $I(b, x_{ab}) = 1$.

SC-graph is much the same as 1-inclusion graph used in (Haussler et al., 1994) to obtain lower bounds on VC-dimension. The VC-dimension may result in highly overestimated generalization bounds as it is based on the union bound. In our combinatorial framework the SC-graph is used to replace the union bound by a much more accurate technique.

Note that SC-graph can be considered also as a subgraph of the Hasse diagram (the graph of transitive reduction) of the partial order over error vectors.

SC-bound for pessimistic Empirical Risk Minimization.

Lemma 7 *If learning algorithm μ is pessimistic ERM, then conjecture 4 holds, and for any $a \in A$*

$$X_a = \bigl\{x_{ab} \in \mathbb{X} \bigm| a \prec b\bigr\} \quad \textit{is the protective subset;}$$
$$X'_a = \bigl\{x \in \mathbb{X} \bigm| \exists b \in A \colon b \leqslant a,\ I(b, x) < I(a, x)\bigr\} \quad \textit{is the prohibitive subset.}$$

Proof: Let us use a proof by contradiction showing that if $\mu X = a$, then $X_a \subseteq X$ and $X'_a \subseteq \bar{X}$.

Assume that an object $x_{ab} \in X_a$ not belonging to X exists. Then $n(a, X) = n(b, X)$ because the error vectors a and b differ by exactly one object x_{ab}. At the same time $n(a, \mathbb{X}) + 1 = n(b, \mathbb{X})$, therefore the learning algorithm μ being pessimistic learns the classifier b rather than a from the training sample X which contradicts the initial condition $\mu X = a$. Then we conclude that $X_a \subseteq X$.

Assume that an object $x \in X'_a$ belonging to X exists. Then $n(b, X) < n(a, X)$. The learning algorithm μ being empirical risk minimizer learns the classifier b rather than a from the training sample X which contradicts the initial condition $\mu X = a$. Then we conclude that $X'_a \subseteq \bar{X}$. ■

Corollary 8 *Any classifier $a \in A$ produces errors on all prohibitive objects X'_a and does not produce errors on all protective objects X_a.*

Upper connectivity $q(a) = |X_a|$ of a classifier a is the *out-degree* of the vertex a in the SC-graph, i. e. the number of edges leaving the vertex a.

Lower connectivity $d(a) = |X'_a|$ of a classifier a is the *in-degree* of the vertex a in the SC-graph, i. e. the number of edges entering the vertex a.

Inferiority $r(a) = |X'_a|$ of a classifier a is the number of different objects assigned to edges below the vertex a in the SC-graph. If a correct classifier $a_0 \in A$ exists such that $n(a_0) = 0$, then inferiority is equal to the number of errors, $r(a) = n(a)$. In general case, $d(a) \leqslant r(a) \leqslant n(a)$.

Theorem 9 (SC-bound) *If learning algorithm μ is ERM, then for any $\varepsilon \in [0, 1]$ the probability of overfitting is bounded by the weighted sum of FC-bounds over the set A:*

$$Q_\varepsilon(\mu, \mathbb{X}) \leqslant \sum_{a \in A} \frac{\binom{L-q-r}{\ell-q}}{\binom{L}{\ell}} H_{L-q-r}^{\ell-q,\, m-r} \left(\tfrac{\ell}{L}(m - \varepsilon k)\right), \tag{7}$$

where $q = q(a)$ is upper connectivity, $r = r(a)$ is inferiority, $m = n(a)$ is the number of errors of classifier a on the general object set $\mathbb{X}$.

Proof: The bound (7) for pessimistic ERM follows immediately from theorem 6, lemma 7, and corollary 8. From lemma 1 it follows that (7) also holds for any ERM. ■

The weight $P_a = \binom{L-q-r}{\ell-q} / \binom{L}{\ell}$ in the sum (7) is an upper bound on the probability to learn the classifier a. Its value decreases exponentially as connectivity $q(a)$ and inferiority $r(a)$ increase. This fact has two important consequences.

First, connected sets of classifiers are less subjected to overfitting. Note that an attempt to use only the fact of connectedness with no counting the number of connections did not lead to a tight bound (Sill, 1998).

Second, only a little part of lower layers contribute significantly to the probability of overfitting. This fact encourages effective procedures for level-wise bottom-up SC-bound computation.

The SC-bound (7) is much more tight than the VC-bound (4). It can be transformed into the VC-bound by substituting $q = r = 0$, i. e. by totally disregarding the SC-graph structure.

SC-bound for Discrepancy Maximization.

Lemma 10 *If μ is discrepancy maximization, then conjecture 4 holds, and for any $a \in A$*

$$X_a = \bigl\{x_{ab} \in \mathbb{X} \bigm| a \prec b\bigr\} \text{ is the protective subset;}$$
$$X'_a = \bigl\{x_{ba} \in \mathbb{X} \bigm| b \prec a\bigr\} \text{ is the prohibitive subset.}$$

Proof: From $\mu X = a$ it follows that

$$\delta(a, X) \geqslant \delta(b, X) \ \text{ for any } b \in A. \tag{8}$$

Consider an arbitrary $x_{ab} \in X_a$. Then $I(a, x_{ab}) = 0,\ I(b, x_{ab}) = 1$. If we assume that $x_{ab} \in \bar{X}$, then $\delta(b, X) = \nu(b, \bar{X}) - \nu(b, X) = \nu(a, \bar{X}) + \frac{1}{k} - \nu(a, X) > \delta(a, X)$ which contradicts (8). Therefore, our assumption is false and $x_{ab} \in X$. Since x_{ab} is arbitrary element of X_a, we have $X_a \subseteq X$.

Consider an arbitrary $x_{ba} \in X'_a$. Then $I(b, x_{ba}) = 0,\ I(a, x_{ba}) = 1$. If we assume that $x_{ba} \in X$, then $\delta(b, X) = \nu(b, \bar{X}) - \nu(b, X) = \nu(a, \bar{X}) - \nu(a, X) + \frac{1}{\ell} > \delta(a, X)$ which contradicts (8). Therefore our assumption is false and $x_{ba} \in \bar{X}$. Since x_{ba} is arbitrary element of X'_a, we have $X'_a \subseteq \bar{X}$. ■

Lemma 10 for discrepancy maximization is similar to lemma 7 for pessimistic ERM. The difference in prohibitive subsets leads to the bound similar to (7), in which inferiority $r(a)$ changes to lower connectivity $d(a)$. We call this bound the *uniform connectivity* (UC) bound for two reasons. First, it also holds for the probability of large uniform deviation $\widetilde{Q}_\varepsilon(A, \mathbb{X})$. Second, it does not use the splitting property of the set A any more.

Theorem 11 (UC-bound) *If μ is discrepancy maximization, then for any $\varepsilon \in [0, 1]$ the probability of overfitting is bounded by the weighted sum of FC-bounds over the set A:*

$$Q_\varepsilon(\mu, \mathbb{X}) = \widetilde{Q}_\varepsilon(A, \mathbb{X}) \leqslant \sum_{a \in A} \frac{\binom{L-q-d}{\ell-q}}{\binom{L}{\ell}} H_{L-q-d}^{\ell-q,\, m-d} \left(\tfrac{\ell}{L}(m - \varepsilon k)\right), \tag{9}$$

where $q = q(a)$ is upper connectivity, $d = d(a)$ is lower connectivity, $m = n(a)$ is the number of errors of classifier a on the general object set $\mathbb{X}$.

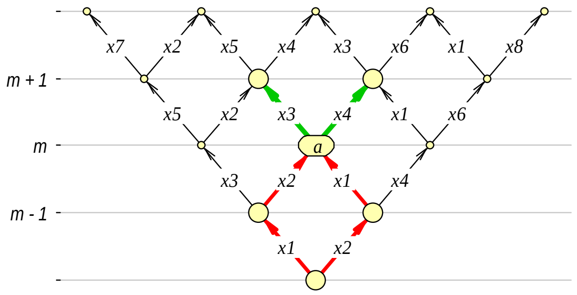


Figure 2: The SC-graph over 2-dimensional monotonic lattice of altitude 4. Algorithm a at m-th layer has prohibitive subset $\{x1, x2\}$ and protective subset $\{x3, x4\}$.

Proof: follows immediately from theorem 6 and lemma 10. ∎

The UC-bound (9) can be much greater (less tight) than the SC-bound (7). The weight $\binom{L-q-d}{\ell-q}/\binom{L}{\ell}$ decreases exponentially as connectivity $q(a)$ or $d(a)$ increases. However it does not decrease as $r(a)$ or $n(a)$ increases. This means that the UC-bound ignores the splitting effect.

6 Exact SC-bounds for model sets of classifiers

The set of classifiers A is called a *model set* if it is defined directly as a set of binary error vectors. Typically, a tight or even exact combinatorial bound on probability of overfitting can be obtained for a model set due to its special "regular" structure. Model sets provide a theoretical understanding of overfitting demonstrating how overfitting may depend on splitting, connectivity, and other properties of the set of classifiers. Exact bounds have been obtained for monotonic and unimodal chain, unity neighborhood of a fixed classifier, pair of classifiers, interval of L-dimensional boolean cube (Vorontsov, 2010), Hamming balls, pencils of monotonic chains, and some other constructions.

Monotonic chain of classifiers can be considered as a model of one-parametric family of classifiers such that continuous moving of a parameter away from its optimal value can only increase the number of errors produced by a classifier on the general sample $\mathbb{X}$. A monotonic chain is defined as a set $A = \{a_0, a_1, \dots, a_H\}$ such that $a_0 \prec a_1 \prec \dots \prec a_H$. This is a simplest set of classifiers with splitting and connectivity. An exact bound on probability of overfitting for pessimistic ERM has been obtained in (Vorontsov, 2010) from the principle of protective and prohibitive subsets:

$$Q_\varepsilon(\mu, \mathbb{X}) = \sum_{t=0}^{\min\{H,k\}} \frac{C_{L-q-t}^{\ell-q}}{C_L^\ell} H_{L-q-t}^{\ell-q,\, m}\left(\tfrac{\ell}{L}(m+t-\varepsilon k)\right), \quad q = [t<H].$$

This bound can also be obtained immediately from the common SC-bound and observation that the SC-graph is a chain, $q(a_t) = [t<H]$, and $r(a_t) = t$ for all $t = 0, \dots, H$.

Multidimensional monotonic lattice of classifiers is a natural multidimensional extension of the monotonic chain. Beside splitting and connectivity it also possesses a dimension. Together these three properties are intrinsic for most real sets of classifiers. An exact bound for this set has been obtained in (Botov, 2011) from the modified layer-wise principle of protective and prohibitive subsets. Here we show that it can be easily obtained from our common SC-bound.

Index vector is a vector of nonnegative integer values $J = (j_1, \dots, j_h)$. Denote $|J| = j_1 + \dots + j_h$. Define a partial order relation on index vectors: $(J \leqslant K) \leftrightarrow (j_d \leqslant k_d$ for all $d = 1, \dots, h)$. Define $(J < K) \leftrightarrow (J \leqslant K$ and $J \neq K)$.

Definition 12 *An h-dimensional monotonic lattice of classifiers of altitude H is a set of classifiers $a_J = \bigl\{a_J \colon |J| \leqslant H\bigr\}$ indexed by h-dimensional index vectors $J = (j_1, \dots, j_h)$ such that*

1) for any index vectors J, K if $J < K$, $|J| \leqslant H$, $|K| \leqslant H$, then $a_J \leqslant a_K$;

2) $n(a_J) = m + |J|$ for some m.

Figure 2 shows an example of 2-dimensional lattice of altitude 4. Like monotonic chain, SC-bound for this model set is exact and easily follows from the common SC-bound.

Lemma 13 *There are only three types of objects in the set $\mathbb{X}$:*
1) m objects on which any classifier gives an error;
2) Hh objects $\bigl\{x_t^d\colon t=0,\dots,H{-}1,\ d=1,\dots,h\bigr\}$ such that $I(a_J,x_t^d)=\bigl[t<j_d\bigr]$ for any $|J|\leqslant H$;
3) other objects on which no classifier gives an error.

The proof is not difficult but needs a number of details, therefore we omit it for reasons of space.

Lemma 14 *If μ is pessimistic ERM, then $\bigl[\mu X{=}a_J\bigr]=\bigl[X_J\subseteq X\bigr]\bigl[X'_J\subseteq\bar X\bigr]$ for all $a_J\in A$, where*

$$\begin{aligned} X_J &= \bigl\{x_t^d\colon t=j_d,\ d=1,\dots,h\bigr\} \ \text{ is the protective subset;}\\ X'_J &= \bigl\{x_t^d\colon t<j_d,\ d=1,\dots,h\bigr\} \ \text{ is the prohibitive subset.}\end{aligned}$$

Proof: Let us prove that if $X_J\subseteq X$ and $X'_J\subseteq\bar X$, then only the classifier a_J can be an output of the learning algorithm: $\mu X=a_J$.

The classifier a_J makes errors only on objects from X'_J all belonging to $\bar X$; then $n(a_J,X)=0$.

Consider a classifier a_K such that $K\nleqslant J$. This means that $k_d>j_d$ for some coordinate d. Then the classifier a_K produces an error on the object $x_{j_d}^d$ that belongs to X_J and hence to the training sample X. Therefore, the learning algorithm μ being empirical risk minimizer will not choose a_K.

Consider a classifier a_K such that $K<J$. Then both a_J and a_K do not make errors on the training sample X. Assume that $k_d<j_d$ for some coordinate d. Then only a_J makes an error on a testing object $x_{k_d}^d$. Therefore learning algorithm μ being pessimistic will not choose a_K.

Thus, only the classifier a_J can be an output of the learning algorithm. ■

Theorem 15 *If the learning algorithm μ is pessimistic ERM, the set A is a monotonic h-dimensional lattice of classifiers with altitude H and $m+Hh\leqslant L$, then*

$$Q_\varepsilon(\mu,\mathbb{X})=\sum_{t=0}^{\min\{H,k\}}\binom{h+t-1}{t}\frac{\binom{L-q-t}{\ell-q}}{\binom{L}{\ell}}H_{L-q-t}^{\ell-q,\,m}\left(\tfrac{\ell}{L}(m+t-\varepsilon k)\right),\quad q=[t{<}H]h.$$

Proof: Let us numerate nonempty layers of the monotonic lattice by $t=0,\dots,H$. Note that $n(a_J)=m+t$ and $t=|J|$ for any classifier $a_J\in A$.

It follows from lemma 13 that $r(a_J)=t$ and $q(a_J)=h$ for all $a_J\in A$ except for classifiers from the last (highest) layer which do not have protective objects: $q(a_J)=0$, $|J|=H$.

All classifiers a_J from one layer $t=|J|$ have equal characteristics q, r, and n. Therefore, the sum over classifiers in (7) can be transformed into the sum over layers $t=0,\dots,H$, with multiplier $|A_{m+t}|=\binom{h+t-1}{t}$ under the sum.

By lemma 14 the inequality (5) transforms into the equality and gives a necessary and sufficient condition for $\mu X=a$. Then, the SC-bound (7) also transforms into the equality. ■

In a particular case, when $h=1$, the bound for monotonic lattice transforms into the bound for monotonic chain. Both theoretical and experimental analysis of monotonic lattices can be found in (Botov, 2011). The purpose of this section is to show that exact bounds can be obtained from the common SC-bound for nontrivial sets of classifiers.

7 Tight SC-bound for threshold conjunctive rule

Rule induction is a well known, deeply studied, and widely used paradigm in machine learning (Rivest, 1987, Cohen and Singer, 1999, Marchand and Sokolova, 2005, Fürnkranz and Flach, 2005, Rückert and Raedt, 2008).

Conjunctive rules. Consider a classification problem with labels $y_i\in\mathbb{Y}$, $i=1,\dots,L$ assigned to each object $x_i\in\mathbb{X}$ respectively. Consider a parametric set R of *conjunctive rules*

$$r(x;\theta)=\prod_{j\in J}\bigl[x^j\leqslant\theta^j\bigr], \tag{10}$$

where $x=(x^1,\dots,x^n)$ is a vector of numerical features of an object x, $J\subseteq\{1,\dots,n\}$ is a subset of features, $\theta^j\in\mathbb{R}$ is a threshold parameter for j-th feature. Note that (10) can be easily generalized to rules with conditions $\theta_1^j\leqslant x^j\leqslant\theta_2^j$ or $x^j\geqslant\theta^j$ simply by adding a negated feature $-x^j$.

An object x is said to be *covered* by the rule r if $r(x)=1$.

Rule learning. The rule induction system usually learns a rule set R_y for each class $y \in \mathbb{Y}$ from a training set X. Two criteria are optimized simultaneously to select useful rules — the number of positive and negative examples covered by r, respectively:

$$p(r, X) = \#\bigl\{x_i \in X \bigm| r(x_i) = 1,\ y_i = y\bigr\} \to \max;$$
$$n(r, X) = \#\bigl\{x_i \in X \bigm| r(x_i) = 1,\ y_i \neq y\bigr\} \to \min.$$

In practice the two-criteria optimization task is reduced to one-criterion task by means of heuristic function $H(p, n)$. Examples of H are entropy, Gini index, Fisher's exact test, χ^2- and ω^2-tests, and many others (Fürnkranz and Flach, 2005).

Rule based classifier. After learning the rule sets R_y for all $y \in \mathbb{Y}$ the classifier can be build up as a composition of rules. The weighted voting is a most popular choice:

$$a(x) = \arg\max_{y \in \mathbb{Y}} \sum_{r \in R_y} w_r r(x),$$

where weights $w_r \geqslant 0$ are to be learned from the training set X. So, there are three things to learn:
1) thresholds $\theta^j,\ j \in J$ for each subset J;
2) feature subset J for each rule r;
3) weight w_r for each rule r.

Respectively, there are three reasons for overfitting. In this work we deal with overfitting resulting from thresholds learning and use the SC-bound to build a new criterion for feature subsets selection. So, our goal is to reduce overfitting for learning stages 1) and 2), with motivation that a good classifier can be hardly build up from overfitted rules. We leave the stage 3) beyond the scope of this paper bearing in mind that the overfitting of weighted voting is now well understood (Schapire et al., 1998, Koltchinskii et al., 2001).

The main idea of heuristic modification is to obtain the SC-bound on both p and n for a fixed J; then to get inverted estimates that hold with probability at least $1 - \eta$:

$$\tfrac{1}{k} p(r, \bar{X}) \geqslant \tfrac{1}{\ell} p(r, X) - \varepsilon_p(\eta),$$
$$\tfrac{1}{k} n(r, \bar{X}) \leqslant \tfrac{1}{\ell} n(r, X) + \varepsilon_n(\eta),$$

and substitute these estimates instead of p, n in a heuristic function:

$$H'(p, n) = H\bigl(p - \ell\varepsilon_p(\eta), n + \ell\varepsilon_n(\eta)\bigr).$$

The modified heuristic H' estimates the quality of a rule with a fixed subset of features J and learnt thresholds. Both discrepancies $\varepsilon_p(\eta)$ and $\varepsilon_n(\eta)$ are very sensitive to the number of rules in bottom levels of the SC-graph. The more rules the bottom levels contain, the higher is overfitting. The modified heuristic H' can be considered as a feature selection criterion based on a data-dependent generalization bound.

In order to specialize the SC-bound for conjunctive rules we first define the binary loss function: $I(r, x_i) = \bigl[r(x_i) \neq [y_i = y]\bigr],\ i = 1, \ldots, L$, for any rule r of class y. Second, we will describe how to iterate in (7) only rules having pairwise different error vectors. As a result we will obtain a heuristic function H modified by the SC-bound which can be easily incorporated into any rule inducer.

Classes of equivalent rules. Fig. 3(a) shows an example of two-dimensional objects set $\mathbb{X}$ combining the scatter plot of initial objects with the set R of rules $r(x; \theta) = [x^1 \leqslant \theta^1]\,[x^2 \leqslant \theta^2]$. Each object and each rule is represented by a node of $(L + 1) \times (L + 1)$ grid in coordinate plane θ^1, θ^2. Nodes containing the rules with equal error vectors are connected. Fig. 3(b) shows the same objects set, but only representative rules (with pairwise different error vectors) are left on the plot. Rules with Hamming distance 1 are connected. Fig. 3(c) shows the SC-graph of the rule set isomorphic to the graph on Fig. 3(b).

From the above example it seems that the equivalence classes of rules have a nontrivial structure. In fact, this is not the case and equivalence classes can be efficiently described and searched.

For the sake of simplicity consider a case when all values $x_i^j,\ i = 1, \ldots, L$ are pairwise different for each feature $j = 1, \ldots, n$. Without loss of generality, assume that the features take integer values $1, \ldots, L$ and the thresholds take integer values $0, \ldots, L$.

The loss function I induces an equivalence relation on the set of rules R. Two rules are equivalent, $r \sim r'$ if their error vectors are equal. Let $u = (u^j)_{j \in J},\ v = (v^j)_{j \in J}$ be a pair of vectors. Define a partial order relation $(u \leqslant v) \leftrightarrow \forall j \in J\ (u^j \leqslant v^j)$. Define $(u < v) \leftrightarrow (u \leqslant v \text{ and } u \neq v)$.

By θ_r^j denote a value of threshold on j-th feature for a rule $r(x; \theta)$, $\theta = (\theta_r^j)_{j \in J}$

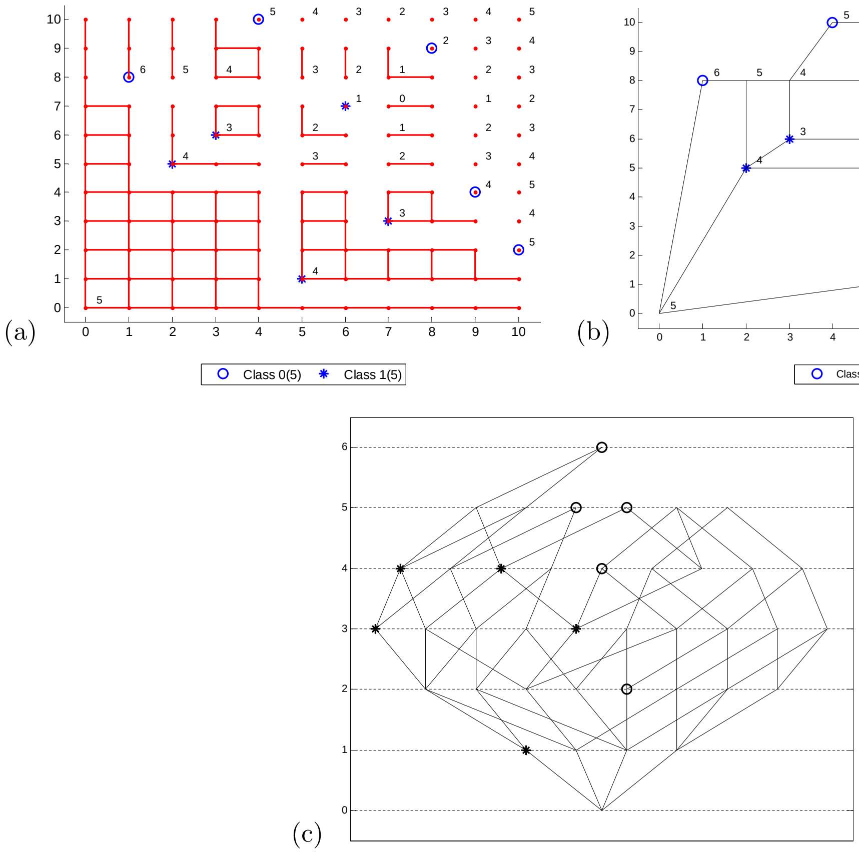


Figure 3: Example of a set of $L = 10$ objects of two classes (a), the graph of representative rules (b), and the SC-graph (c).

Lemma 16 *If $E \subseteq R$ is equivalence class of rules, then a rule $r(x;\theta_E)$ belongs to the equivalence class E, where $\theta_E^j = \min\limits_{r\in E} \theta_r^j$.*

Proof: All equivalent rules $r \in E$ have equal binary error vectors. Then a binary function

$$r_E(x) = \prod_{r\in E} r(x;\theta_r)$$

takes the same value on each object $x \in \mathbb{X}$ as any rule r from E, $r_E(x) = r(x;\theta_r)$. Moreover, the binary function r_E can be represented in the form (10), so it is also a rule:

$$r_E(x) = \prod_{r\in E}\prod_{j\in J}[x^j \leqslant \theta_r^j] = \prod_{j\in J}[x^j \leqslant \min_{r\in E}\theta_r^j] = \prod_{j\in J}[x^j \leqslant \theta_E^j] = r(x;\theta_E).$$

This means that the rule $r(x;\theta_E)$ pertains to the equivalence class E. ■

The rule $r_E(x) \equiv r(x;\theta_E)$ from the previous lemma will be called a *representative rule* of the equivalence class E. On figure 3 representative rules correspond to lower left corners of equivalence classes: $(0,0)$, $(1,8)$, $(2,5)$, $(5,1)$, etc.

A *boundary point* of the subset $S \subseteq \mathbb{X}$ is a vector θ_S with coordinates $\theta_S^j = \max\limits_{x\in S} x^j$, $j \in J$.

A *boundary object* of the subset $S \subseteq \mathbb{X}$ is any object $x \in S$ such that $x^j = \theta_S^j$ for some j.

A *boundary subset* is a subset $S \subseteq \mathbb{X}$ such that all objects $x \in S$ are boundary objects of S.

Empty set is assumed to be a boundary subset with boundary point $\theta_\varnothing^j = 0$, $j \in J$.

Note that $r(x, \theta_S) = 1$ for any $x \in S$.

Lemma 17 *If a threshold vector θ is a boundary point of some boundary subset, then this subset can be unambiguously determined by the threshold vector θ:*

$$S_\theta = \bigcup_{j\in J} \bigl\{x \in \mathbb{X} \mid x^j = \theta^j,\ r(x,\theta) = 1\bigr\}. \tag{11}$$

Proof: If $\theta = \theta_\varnothing$, then $S_\theta = \varnothing$ by definition. Otherwise, threshold vector θ is a boundary point, and for any index $j \in J$ an object $x \in \mathbb{X}$ exists such that $x^j = \theta^j$ and $x^{\bar{\jmath}} \leqslant \theta^{\bar{\jmath}}$ for all other indices $\bar{\jmath} \in J \setminus \{j\}$. Then $r(x,\theta) = 1$, and it follows from (11) that $x \in S_\theta$. The index $j \in J$ was chosen arbitrarily. Therefore, the threshold vector θ is a boundary point of the subset S_θ. The subset S_θ is a boundary because each of its objects is a boundary according to (11). ■

Theorem 18 *Each equivalence class E bijectively corresponds to a boundary subset S: $\theta_E = \theta_S$.*

Proof: Consider an arbitrary equivalence class E with representative rule $r(x;\theta_E)$. By lemma 16 the decrease of threshold θ^j_E along any coordinate $j \in J$ changes the value $r(x;\theta_E)$ on some object $x \in \mathbb{X}$ such that $x \leqslant \theta_E$. This value may only be decreased: $[x^j \leqslant \theta^j_E] = 1$, $[x^j \leqslant \theta^j_E - 1] = 0$. From this it follows that $x^j = \theta^j_E$, then x is a boundary object in the subset $\bigl\{x' \in \mathbb{X}\colon x' \leqslant \theta_E\bigr\}$. As feature j is arbitrary, this means that threshold vector θ_E is a boundary point. According to lemma 17 it determines a unique boundary subset.

The converse is also true. Any boundary subset S has a boundary point θ_S. The rule $r(x;\theta_S)$ is a representative rule of the equivalence class E to which it belongs, because the decrease of the threshold θ^j_S along any coordinate $j \in J$ changes the value $r(x;\theta_S)$ on some boundary object of the subset S.

Thus we conclude that each equivalence class E bijectively corresponds to a boundary subset S such that $\theta_E = \theta_S$. ■

Denote M_q the set of all boundary subsets of size q and recapitulate its properties:

1) M_1 consists of all L one-element subsets, $M_1 = \bigl\{\{x_1\},\dots,\{x_L\}\bigr\}$;
2) M_2 consists of all pairs of incomparable objects from $\mathbb{X}$;
3) the size of boundary subset can not exceed the rank of conjunction: $M_q = \varnothing$ for all $q > |J|$;
4) any boundary subset consists of pairwise incomparable objects; the converse is not obligatory true: a subset of three or more pairwise incomparable objects may not be a boundary subset;
5) any subset $S' \subset S$ of the boundary subset S is a boundary subset too.

A simple algorithm which iterates all boundary subsets follows from these properties. At the first step a subset M_1 is formed from all L one-object subsets. At each subsequent step $q = 2,\dots,|J|$ for all subsets $S' \in M_{q-1}$ and for all objects $x \in \mathbb{X} \setminus S'$ if the subset $S = S' \cup \{x\}$ is boundary (which can be easily tested by definition), then S joins M_q.

More efficient but more complicated level-wise algorithm for SC-bound calculation is described in (Vorontsov and Ivahnenko, 2011). It iterates rules from bottom to upper levels and uses early stopping to bypass rules that do not make significant contribution to the SC-bound. Additionally this algorithm calculates the connectivity and the inferiority of each rule.

Experiment. We use state-of-the art algorithms C4.5 (Quinlan, 1996), C5.0 (Quinlan, 1993), RIPPER (Cohen, 1995), and SLIPPER (Cohen and Singer, 1999) as baseline rule learners. Our rule learning engine is based on breadth-first search as features selection strategy. Fisher's exact test (Martin, 1997) is used as heuristic H. To build compositions of rules we use three algorithms. Logistic Regression (LR) is a linear classifier that aggregates rules learned independently. Weighted Voting (WV) is a boosting-like ensemble of rules, similar to SLIPPER, which trains each next rule on reweighted training set. Decision List (DL) is a greedy algorithm, which trains each next rule on training objects not covered by all previous rules.

There are two modifications of heuristic $H'(p,n)$. The SC-modification uses SC-bound on the probability of overfitting Q_ε as described above. The MC-modification uses the Monte-Carlo estimation of Q_ε via 100 random partitions $\mathbb{X} = X \sqcup \bar{X}$. For both modifications we set $\ell = k$.

Table 1 shows that initially our algorithms WV, DL are comparable to the baseline. WV outperforms DL, which corresponds to the results of other authors. Both SC- and MC- modifications reduce overfitting significantly and always outperform their respective initial versions. The difference between SC- and MC- modifications is not significant. Then, a moderate looseness of the SC-bound does not reduce its practical usefulness as a rule selection criterion.

Table 1: Experimental results on 6 real data sets from UCI Machine Learning Repository. For each pair ⟨task, algorithm⟩ an average testing error obtained from 10-fold cross validation is given, in percents. For each task three best results are bold-emphasized. Algorithms 1–7 are baseline rule learners. Our algorithms: WV — Weighted Voting, DL — Decision List, SC — using heuristic modified by SC-bound, MC — using heuristic modified by Monte-Carlo estimation of overfitting.

	algorithms	tasks					
		australian	echo-card	heart dis.	hepatitis	labor	liver
1	RIPPER−opt	15.5	**2.9**	19.7	20.7	18.0	32.7
2	RIPPER+opt	15.2	5.5	20.1	23.2	18.0	**31.3**
3	C4.5 (Tree)	14.2	5.5	20.8	18.8	14.7	37.7
4	C4.5 (Rules)	15.5	6.8	20.0	18.8	14.7	37.5
5	C5.0	**14.0**	4.3	21.8	20.1	18.4	31.9
6	SLIPPER	15.7	4.3	**19.4**	**17.4**	**12.3**	32.2
7	LR	14.8	4.3	19.9	18.8	14.2	32.0
8	WV	14.9	4.3	20.1	19.0	14.0	32.3
9	DL	15.1	4.5	20.5	19.5	14.7	35.8
10	**WV+MC**	**13.9**	**3.0**	**19.5**	**18.3**	**13.2**	**30.7**
11	**DL+MC**	14.5	3.5	19.8	18.7	13.8	32.8
12	**WV+SC**	**14.1**	**3.2**	**19.3**	**18.1**	**13.4**	**30.2**
13	**DL+SC**	14.4	3.6	19.5	18.6	13.6	32.3

8 Conclusion

Splitting and connectivity (SC) are very important data-dependent properties of a set of classifiers that determines its generalization ability. This work gives a combinatorial SC-bound on the probability of overfitting. The SC-bound takes into account both the learning algorithm and a fine internal structure of the set of classifiers represented by the SC-graph. The SC-bound can be exact (non-asymptotical, non-overestimated) for some nontrivial sets of classifiers like the monotonic chain and the multidimensional monotonic lattice.

If the discrepancy maximization is considered as a learning algorithm, then the SC-bound transforms into the UC (uniform connectivity) bound and gives the probability of a large uniform deviation. Thus, the UC-bound can be used for the Rademacher complexity estimation. Note that the UC-bound takes into account the connectivity but neglect the splitting.

The SC-bound being applied to threshold conjunctive rules helps to select features for each rule more accurately, and then to reduce overfitting of a rule induction machine. In practice this can be implemented as a slight modification of a heuristic function which estimates the usefulness of a rule. Experiments on six real data sets show that the proposed modification reduces overfitting.

References

P. V. Botov. Exact bounds on probability of overfitting for multidimensional model sets of classifiers (to appear). *Pattern Recognition and Image Analysis*, 2011.

S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16(3):277–292, 2000.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, (9):323–375, 2005.

W. W. Cohen. Fast effective rule induction. In *Proc. of the 12th International Conference on Machine Learning, Tahoe City, CA*, pages 115–123. Morgan Kaufmann, 1995.

W. W. Cohen and Y. Singer. A simple, fast and effective rule learner. In *Proc. of the 16 National Conference on Artificial Intelligence*, pages 335–342, 1999.

J. Fürnkranz and P. A. Flach. Roc 'n' rule learning-towards a better understanding of covering algorithms. *Machine Learning*, 58(1):39–77, 2005.

R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics.* Reading, Massachusetts: Addison-Wesley, 1994.

D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Inf. Comput.*, 115:248–292, December 1994.

V. Koltchinskii, D. Panchenko, and F. Lozano. Further explanation of the effectiveness of voting methods: The game between margins and weights. In *14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001, Amsterdam, The Netherlands, July 2001, Proceedings*, volume 2111, pages 241–255. Springer, Berlin, 2001.

J. Langford. *Quantitatively Tight Sample Complexity Bounds.* PhD thesis, Carnegie Mellon Thesis, 2002.

G. Lugosi. On concentration-of-measure inequalities. Machine Learning Summer School, Australian National University, Canberra, 2003.

M. Marchand and M. Sokolova. Learning with decision lists of data-dependent features. *Journal of Machine Learning Reasearch*, 6:427–451, 2005.

J. K. Martin. An exact probability metric for decision tree splitting and stopping. *Machine Learning*, 28(2-3):257–291, 1997.

P. Philips. *Data-Dependent Analysis of Learning Algorithms.* PhD thesis, The Australian National University, Canberra, 2005.

J. R. Quinlan. *C4.5: Programs for machine learning.* Morgan Kaufmann, San Francisco, CA, 1993.

J. R. Quinlan. Bagging, boosting, and C4.5. In *AAAI/IAAI, Vol. 1*, pages 725–730, 1996.

R. L. Rivest. Learning decision lists. *Machine Learning*, 2(3):229–246, 1987.

U. Rückert and L. De Raedt. An experimental evaluation of simplicity in rule learning. *Artif. Intell.*, 172(1):19–28, 2008.

R. E. Schapire, Y. Freund, We Sun Lee, and P. Bartlett. Boosting the margin: a new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26(5):1651–1686, 1998.

J. Sill. *Monotonicity and connectedness in learning systems.* PhD thesis, California Institute of Technology, 1998.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

V. N. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

N. Vayatis and R. Azencott. Distribution-dependent Vapnik-Chervonenkis bounds. *Lecture Notes in Computer Science*, 1572:230–240, 1999.

K. V. Vorontsov. Combinatorial probability and the tightness of generalization bounds. *Pattern Recognition and Image Analysis*, 18(2):243–259, 2008.

K. V. Vorontsov. Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting. *Pattern Recognition and Image Analysis*, 19(3):412–420, 2009.

K. V. Vorontsov. Exact combinatorial bounds on the probability of overfitting for empirical risk minimization. *Pattern Recognition and Image Analysis*, 20(3):269–285, 2010.

K. V. Vorontsov and A. A. Ivahnenko. Tight combinatorial generalization bounds for threshold conjunction rules (to appear). In *4th International Conference on Pattern Recognition and Machine Intelligence (PReMI'11). June 27 – July 1, 2011*, 2011.